MEMORANDUM REPORT ARBRL-MR-03251

AN ALGORITHM FOR COMPUTING DERIVATIVES OF NOISY DATA

Ceslovas Masaitis
George Francis

March 1983

US ARMY ARMAMENT RESEARCH AND DEVELOPMENT COMMAND
BALLISTIC RESEARCH LABORATORY
ABERDEEN PROVING GROUND, MARYLAND

Destroy this report when it is no longer needed.
Do not return it to the originator.

Additional copies of this report may be obtained from the National Technical Information Service, U. S. Department of Commerce, Springfield, Virginia 22161.

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

The use of trade names or manufacturers' names in this report does not constitute indorsement of any commercial product.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER MEMORANDUM REPORT ARBRL-MR-03251	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* AN ALGORITHM FOR COMPUTING DERIVATIVES OF NOISY DATA		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR*(s)* C. Masaitis and G. Francis		8. CONTRACT OR GRANT NUMBER*(s)*
9. PERFORMING ORGANIZATION NAME AND ADDRESS US Army Ballistic Research Laboratory ATTN: DRDAR-BLB Aberdeen Proving Ground, MD 21005		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 1L162618AH80
11. CONTROLLING OFFICE NAME AND ADDRESS US Army Armament Research & Development Command US Army Ballistic Research Laboratory (DRDAR-BL) Aberdeen Proving Ground, MD 21005		12. REPORT DATE March 1983 13. NUMBER OF PAGES 28
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Approximation
Numerical differentiation
Estimation of derivatives
Smoothing
Curve fitting
Differentiation of noisy data
Extrapolation

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

An algorithm for computing derivatives of noisy data, as described by the authors in two previous reports, is modified, and empirical support for the choice of relative weights for approximating functions is provided.

The algorithm is applied to ten sets of synthetic data generated by evaluating analytic functions and addition pseudo-random errors to their values. The accuracy of the estimates of the first and second derivatives is on the average (contd)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

(Block 20 continued)

almost five times better than that obtained by cubic B-spline approximation. Extrapolation is considered.

TABLE OF CONTENTS

LIST OF TABLES

Page

1. INTRODUCTION

Preceding reports [1,2] by the authors describe an algorithm for computing derivatives of a function of a real variable from its tabular data. The algorithm is based on two sets of assumptions. The first set specifies a class of functions such that the tabular data are approximate values of one of these functions. The second set of assumptions determines a finite dimensional subspace A_k of the selected functions from which an approximation of the tabular data is selected. The essential difference between the current algorithm (i.e., the algorithm described in this report) and any of the standard procedures such as moving polynomial arc, linear regression, spline approximation, and others is due to this second set of assumptions. The standard methods assume directly or indirectly that an element of a preselected finite dimensional space, say S_k, yields an appropriate approximation and that S_k is independent of the data at hand. The "best" approximation in S_k is obtained by applying a preselected criterion characteristic of the algorithm. In contrast to this our procedure examines the data and selects a linear subspace A_k of an assumed infinite dimensional algebra A that is most appropriate for the available data and then determines a family of approximations in the subspace A_k dependent on the data and also the relative accuracy of these approximations expressed by weights which in their turn depend on the data. Final results are weighted averages of individual approximations.

Usually a subspace S_k and its approximating element are selected by a heuristic criterion such as minimum root mean square error, degree of smoothness, and others. In most cases this criterion represents a compromise between simplicity (smoothness, dimension of S_k) of an approximating function and its faithfulness to the data. Similarly, our procedure selects approximating elements and their weights by a compromise between the accuracy of approximation and the robustness of the approximating model. The criterion for this compromise is purely heuristic. In this report we examine such a criterion, which is a modification of one used in the previous reports. [1,2]

We note that approximating functions selected by the algorithm[2] yield no values of derivatives in the initial and final segments of the data, where the lengths of these segments are dependent on the data. The number of points lost in this way depends on the dimension k of the subspace A_k and on the selected multiple q of the data step size. Here we describe a procedure for computing derivatives at the beginning and the end of the data sequence. These are obtained with the aid of digital linear filters dependent on the data.

---

[1] *C. Masaitis and G. Francis, "Numerical Differentiation of Noisy Data," ARBRL-MR-03126, Aug 81, AD A104631.*

[2] *C. Masaitis and G. Francis, "Computation of Derivatives from Tabular Data," ARBRL-MR-03188, Jul 82, AD B066124L.*

2. ALGORITHM

Details of the algorithm for computing derivatives are described in the report [1] which is denoted by ND in the references below. A summary of these details follows.

Various formulas of this algorithm are derived from four assumptions:

a. Tabular data $x(n), n = 1,2,..,N$ are obtained by measuring a function $y(t)$ at equally spaced points with a step size Δ. The measuring error ϵ_n is a white noise with zero mean and unknown variance σ^2 independent of time, i.e. $x(n) = y(n\Delta) + \epsilon_n$.

b. The function $y(t)$ belongs to a family of functions whose derivatives on the interval $[0,T]$ can be represented by a linear combination of values of $y(t)$, i.e.

$$y'(t) = \sum_{i=1}^{k} a_i(v) y[t + (v - i)\Delta] \tag{2.1}$$

for any integer v satisfying $-t + k\Delta \leqslant v\Delta \leqslant T - t$. The coefficients $a_i(v)$ depend on the function $y(t)$ and the location parameter v but not on t.

As shown in Section 3 of *ND* assumption (b) implies that $y(t)$ belongs to the algebra A generated by algebraic, trigonometric, and exponential polynomials defined on the interval $[0,T]$.

According to Section 1 of *ND* an assumption that $y(t) \in A$ is equivalent to an assumption that for every $\Delta > 0$ $y(t)$ satisfies the difference equation:

$$P_\Delta(B_\Delta)\, y(t) = 0, \tag{2.2}$$

where $P_\Delta(\lambda) = \prod_{j=1}^{m} (\lambda - \lambda_j^\Delta)^{n_j+1}$ and $B_\Delta\, y(t) = y(t - \Delta)$. Hence, an approximating function can be obtained by constructing an autoregressive model such as (2.2).

We write $x(n,p,q)$ for $x(p + qn)$, for any positive integer q and $p = 0, 1, 2,...,q - 1, n = 0,1,...,N_p$, where $N_p = [\frac{N-p}{q}]$, i.e.

$$x(n,p,q) = y(r\Delta) + \epsilon_r, \tag{2.3}$$

where $r = p + qn$. By substituting this in (2.2) we get

$$x(n,p,q) - \epsilon_r = \sum_{j=1}^{k} a_j \left[x(n-j,p,q) - \epsilon_{r-j} \right]. \tag{2.4}$$

According to Section 2 of ND the minimum variance estimates of the a_j's are obtained by the following iterative procedure.

Let $M\ (k,q)$ be the matrix of the normal equations of the overdetermined system

$$\sum_{j=1}^{k} a_j\, x\ (n-j,\ p,\ q) = x\ (n,\ p,\ q), \tag{2.5}$$

$n = k+1,\ldots, N_p, p = 0, 1,\ldots, q-1$. Let $X\ (k,q)$ be the right-hand side vector of these normal equations and $\tilde{N} = \sum_{p=0}^{q-1} (N_p - k)$. Let

$$\sigma_u{}^2\ (k,\ q) = \frac{\frac{1}{\tilde{N}} \sum_{p=0}^{q-1} \sum_{n=k+1}^{N_p} \left[x\ (n,\ p,\ q) - \sum_{j=1}^{k} a_j^{(u)} x\ (n-j,\ p,\ q) \right]^2}{1 + \sum_{j=1}^{k} \left[a_j^{(u)} \right]^2} \tag{2.6}$$

for $u = 1, 2,\ldots$ and $\sigma_0^2 = 0$. Let $a^{(u)}$ be the column vector whose components are the approximate values of the coefficient, a_j 's, obtained on the u-th iteration by solving the system of linear equations

$$(M(k,\ q) - \tilde{N}\ \sigma_{u-1}^2\ (k,\ q)\ I)\ a^{(u)} = X(k,q), \tag{2.7}$$

where I is an identity matrix. In summary, we choose positive integers k and q and use the tabular data $x\ (n)$ to generate the augmented matrix $(M(k,\ q), X(k,\ q))$ corresponding to the overdetermined system (2.5). Then by starting with $\sigma_0^2 = 0$ we obtain the first approximations, $a_j^{(1)}$'s, by solving (2.7). These values are substituted in (2.6) to obtain σ_1^2 which in turn is used in (2.7) to obtain the second approximation vector $a^{(2)}$, and so on. The iteration is continued until $\mid a_j^{(u+1)} - a_j^{(u)} \mid < \delta = 10^{-6}$ or $u = 20$. Together with a solution of (2.7) the determinant of this system is also computed. The absolute value of the determinant of the last iteration is denoted by $D\ (k,q)$. Thus, at the end of the iteration we have

$$\sigma^2(k,q) \tag{2.8}$$

computed by (2.6),

$$(a_1\ (k,q),\ a_2\ (k,q),\ldots,\ a_k\ (k,q)), \tag{2.9}$$

and

$$D\ (k,q). \tag{2.10}$$

The corresponding autoregressive model is

$$x\ (n,p,q) = \sum_{j=1}^{k} a_j\ (k,q)\ x\ (n-j,\ p,q). \tag{2.11}$$

The eigenvalues of the autoregressive model, namely

$$\lambda_1(k,q),\ \lambda_2(k,q),\ldots,\ \lambda_k(k,q), \tag{2.12}$$

are the roots of the equation

$$\lambda^k(k,q) = \sum_{j=1}^{k} a_j(k,q)\,\lambda^{k-j}(k,q). \tag{2.13}$$

It is very unlikely that the estimates of the a_j's obtained from the tabular data will yield multiple roots of (2.13). Therefore, we assume here that all the roots are distinct, since all the cases of synthetic data examined in this report and empirical data of the previous report [2] produced no pairs of nearly equal roots. We note that exact representation of $y(t) = 4^t(1 + t + t^2 + t^3)$ used below to generate one set of synthetic data does yield four equal roots. However, an approximation of the values of this function with an added noise does not produce an autoregressive model with equal or nearly equal eigenvalues.

As stated above, an approximating function is an element of the algebra A. Hence, with simple eigenvalues we take

$$x(n,p,q) = \sum_{j=1}^{k} c_j \lambda_j^n \tag{2.14}$$

for all data points with the same value of n. The coefficients, c_j's, are selected to minimize the RMSE of the resulting approximation over a span, say, from $n-K$ to $n+K$. We select $K = k$. Thus, the coefficients, c_j's, are obtained by solving the normal equations of the overdetermined system (2.14), i.e., by solving the following:

$$\sum_{u=-K}^{K} \sum_{j=1}^{k} c_j \lambda_j^{n+u} \lambda_i^{n+u} = \sum_{u=-K}^{K} x(n+u,p,q)\,\lambda_i^{n+u}, i = 1,2,\ldots,k. \tag{2.15}$$

Let $C_k^T = (c_1, c_2, \ldots, c_k)$, $s_{ij} = \sum_{u=-k}^{k} \lambda_i^u \lambda_j^u$, S be the matrix (s_{ij}), and Z be the column vector with components $\sum_{u=-k}^{k} \lambda_j^u x(n+u,p,q)$, $j = 1,2,\ldots,k$. Then (2.15) can be rewritten in equivalent matrix form: $S\,C_k = Z$. Consequently,

$$C_k = S^{-1} Z \tag{2.16}$$

Let Y_n be the column vector with components $x(n+u,p,q)\lambda_j^u$, $j = 1,2,\ldots,k$. Then we get from (2.16):

$$C_k = \sum_{u=-k}^{k} S^{-1} Y_u. \tag{2.17}$$

By writing $\Lambda^T = (\lambda_1^u, \lambda_2^u, \ldots, \lambda_k^u)$ and by substituting (2.17) in (2.14) we get the smoothed value of the function:

$$\hat{x}_k(n, p, q) = \sum_{u=-k}^{k} \Lambda^T S^{-1} Y_u. \tag{2.18}$$

In view of (2.3) the relation (2.14) can be written as follows: $\hat{x}_k(t) = \sum_{j=1}^{k} c_j \lambda_j^{q\frac{t}{\Delta}-\frac{p}{q}}$, or

$$\hat{x}_k(t + p\Delta) = \sum_{j=1}^{k} c_j \lambda_j^{q\frac{t}{\Delta}} . \tag{2.19}$$

Differentiation of (2.19) s times yields

$$\frac{d^s \hat{x}(t + p\,\Delta)}{d\,t^s} = \frac{1}{(q\,\Delta)^s} \sum_{j=1}^{k} c_j \lambda_j^{q\frac{t}{\Delta}} \log^s \lambda_j . \tag{2.20}$$

We replace t in (2.20) by $t - p\Delta$, and then we let $t = r\,\Delta$ with r as in (2.3). This in turn leads to

$$\hat{x}_k^{(s)}(n,p,q) = \frac{1}{(q\,\Delta)^s} \sum_{j=1}^{k} c_j \lambda_j^n \log^s \lambda_j . \tag{2.21}$$

In view of (2.17) the estimate of the s—th derivative in (2.21) can be written as follows:

$$\hat{x}_k^{(s)}(n,p,q) = \frac{1}{(q\Delta)^s} \sum_{u=-k}^{k} \Lambda_s^T S^{-1} Y_u \tag{2.22}$$

where $\Lambda_s = d^s \Lambda / du^s$ is a column vector with components $\lambda_j^u \log^s \lambda_j, j = 1,2,...,k$ This completes the description of the basic relations of the algorithm.

3. RELATIVE ACCURACY OF APPROXIMATIONS

The algorithm of the preceding section yields several approximating functions corresponding to various choices of k and q. According to the choices in an earlier report[2] there are $3\left[\min(\frac{N}{17},39)\right]$ approximating functions for a data set of N points. Here the square brackets denote an integer not exceeding the expression in the brackets. We discuss now a heuristic rule to assign a weight to each approximation.

As described in Section 2 of ND all the approximations corresponding to autoregressive models with eigenvalues containing a real negative part are assigned weight zero, since in this case an approximating function contains a periodic term with the period so short that less than four data points are contained in the period; i.e., the data are inadequate to determine this term with sufficient accuracy. We illustrate this by an approximation to $x(t) = \text{erf}(t) + \epsilon$, where ϵ is a normally distributed random variable with standard deviation σ. We generate data by evaluating $x(t)$ in the interval [0,2] with the step size $\Delta = 0.005$ and by adding pseudorandom numbers ϵ. Autoregressive models computed for the values of this function plus random error with $\sigma = 0.00119$ corresponding to eigenvalues with

positive real parts allow a maximum frequency $\omega_{max} = 0.45\pi$. If only the terms with the frequencies $\omega_i = \frac{i\pi}{N}$ satisfying the condition $\omega_{i-1} < \omega_{max} \leq \omega_i$ are included in a truncated Fourier series we obtain an approximation with RMSE's equal to 0.087, 0.420, and 1.10 for the function, first, and second derivatives, respectively. If the terms of the next higher frequency are added to this approximation, the RMSE's are 0.068, 0.537 and 2.41. Thus, we have a slight improvement in the values of the function, as expected, but more than twice as big an error in the second derivative. This example shows that rejection of the autoregressive models with negative parts in their eigenvalues is appropriate.

We made similar tests in two additional cases, namely, with the data obtained by evaluating the Bessel function $J_o(t)$ of the first kind (500 points in the interval [1, 6] with one-sigma error of 0.00344) and by evaluating $\sin 2\pi t + 0.1\sin 10\pi t$ (250 points in the interval [0, 1] with one-sigma error of 0.022).

In the first case, if only the terms with the frequencies $\omega_j \leq \omega_i$ where $\omega_{i-1} < \omega_{max} \leq \omega_i$ are included in the Fourier expansion the corresponding errors are 0.121, 0.255, and 0.334. If the terms with frequency ω_{i+1} are added the errors are 0.090, 0.296, and 0.588. Again we have a slight improvement in the values of the function, but derivatives become less accurate.

In the second case the terms with $\omega_j, j = 1,2,3,4,5$ are allowed in the Fourier expansion, according to the same criterion. The resulting RMSE's in the values of the function, first, and second derivatives are 0.0038, 0.074, and 1.73. If the terms with the frequency ω_5 are dropped, the RMSE's become 0.07013, 2.23, and 69.6, respectively. If the terms with $\omega_j, j = 1,2,3,4,5,6$ are included the errors are 0.01169, 0.209, and 6.09. Thus, here again the selected criterion for cut-off frequency in the representation of the data yields optimal results.

Similarly, comparison of approximations computed by the method of the preceding section corresponding to autoregressive models with negative real parts and those corresponding to positive real parts shows that rejection of approximations with negative real parts in their eigenvalues is appropriate.

For instance, for $x(t) = \sin 2\pi t$ (with noise) evaluated in [0, 1] at a step of 0.004 the constraints imposed on k and q allow a total of 48 models. Of these, 24 contain eigenvalues with negative real parts. The most accurate approximation among these yields RMSE's of 0.069 and 2.17 in the first and second derivatives, while the best approximation with positive real parts in eigenvalues yields 0.0397 and 0.0920, respectively.

The remarks in Section 2 of ND and the examples just presented provide heuristic and empirical justification for rejecting (i.e., assigning weights of zero to) the approximations corresponding to autoregressive models with negative real parts in their eigenvalues.

The weights assigned in a previous report[2] to the approximations with positive real parts in their eigenvalues produced relatively good results. Here we examine a modified form of the weights that yields somewhat better results. We assume as before that the weight increases as the RMSE $\sigma(k,q)$ of the solution of the overdetermined system

(2.5) decreases and that it also increases as the corresponding normal system becomes less sensitive to perturbations of the data, i.e., as the absolute value of det $M(k,q)$ in (2.10) increases. We select a simple function satisfying these conditions and invariant with respect to rescaling of the dependent variable x; i.e., we choose the weight $w(k,q)$ given by

$$w(k,q) = \frac{D(k,q)}{\sigma^k(k,q)} \tag{3.1}$$

Comparison of accuracy of various models provides empirical support for this choice of their relative weights. Eight sets of synthetic data yield a total of 196 approximating functions for various values of k and q with non-negative real parts of eigenvalues. The number of the models of this kind for individual data sets range between 16 and 37. Thus, we have $196 \times 2 = 392$ approximations of the first and second derivatives. It turns out that the three highest weights computed by (3.1) for each set of data differ very little, and the values of the weights drop rapidly for the remaining models. For this reason we consider the three models with the highest weights for each set of data. We have altogether 48 approximations of the first and second derivatives with the weights among the three highest. Among these there are 32 approximations that produce smaller errors in the respective derivatives than the remaining 360 approximations. Nine out of sixteen approximations with the highest weights in their respective sets actually are best for the first or second derivative. The remaining seven models with highest values of $w(k,q)$ do not yield best approximations. However, they are not much inferior to the best models of the autoregressive type. In fact, the average reduction of the RMSE of derivatives between the models with the highest $w(k,q)$ and the best ones by comparison with exact values of derivatives is only 32% for these seven cases. The results of the next section show that the RMSE of derivatives computed by a spline approximation is higher in one case (second derivative of e^t) by as much as 26700% than the RMSE obtained by the current method. Even if this extreme case is ignored, the RMSE of the approximations computed by the spline procedure is 474% higher than that of the current method. In view of this a 32% increase in those cases in which the criterion (3.1) fails to select the best model is not significant. Consequently, we accept this criterion on the basis of empirical results.

The smoothed functional values and estimates of derivatives for each (k,q) pair are given by (2.18) and (2.22), respectively, with $p = 0,1,2,\cdots,q-1$ and $n = k+1, k+2, \cdots, N_p - k$, where N_p is the same as in (2.6). This produces no smoothed values for $t = \Delta, 2\Delta, \cdots, kq\Delta$, and for $t = (N - kq + 1)\Delta, (N - kq + 2)\Delta, \cdots, N\Delta$. Denote the values of k and q corresponding to the three models with highest values of $w(k,q)$ by $(k_i, q_i), i = 1,2,3$ and let $n_o = \max(k_i q_i)$. We obtain the smoothed values and approximation of derivatives based on all three models with highest weights only at the points $t = (n_o + 1)\Delta, (n_o - 2)\Delta, \cdots, (N - n_o)\Delta$. These points belong to what we call here the central interval of the data. The smoothed functional values $\hat{x}(n)$ and approximate derivatives $\hat{x}^{(s)}(n)$ at the points of the central interval are defined by the following:

$$\hat{x}(r) = \sum_{k,q} w(k,q)\, \hat{x}_k(n,p,q) \Big/ \sum_{k,q} w(k,q) \tag{3.2}$$

and

$$\hat{x}^{(s)}(r) = \sum_{k,q} w(k,q)\, \hat{x}^{(s)}(n,p,q) \Big/ \sum_{k,q} w(k,q) \tag{3.3}$$

with $r = p + qn$.

Instead of estimating derivatives at the data points not included in the central interval with the aid of available models of lesser accuracy (such as with $k = q = 1$) we adopt the filtering procedure described in the next section.

4. DERIVATIVES IN THE INITIAL AND FINAL SEGMENTS OF THE DATA

Let $k_o = \max(k_1, k_2, k_3)$, where k_i, $i = 1,2,3$ are the parameters of the three models with the highest weights $w(k_i, q_i)$, and let q_o be the value of q_i, such that $k_i = k_o$. Equations (3.2) and (3.3) yield $\hat{x}(r)$ and $\hat{x}^{(s)}(r)$ for $r = (n_o + 1)\Delta$, $(n_o + 2)\Delta, \cdots, (n - n_o)\Delta$. We assume digital filtering models:

$$\hat{x}'(j) = \sum_{i=o}^{p_o} d_i \hat{x}(j - iq_o) + \sum_{i=1}^{r_o} f_i\, \hat{x}'(j - iq_o) \tag{4.1}$$

for each pair of non-negative integers p_o and r_o satisfying the conditions $p_o \leqslant k_o$, $r_o \leqslant k_o$.

If j is such that $n_o + q_o \max(p_o, r_o) \leqslant j \leqslant N - n_o$ then the values $\hat{x}'(j)$, $\hat{x}(j - iq_o)$ and $\hat{x}'(j - iq_o)$ belong to the central interval of the data. Hence, for these values of j (4.1) constitutes an overdetermined system with unknowns d_i and f_i. We solve this system by the least squares method and obtain the estimates $d_i(r_o, p_o)$ and $f_i(r_o, p_o)$, the RMSE $\sigma(p_o, r_o)$ of the solution, and the absolute value $E(p_o, r_o)$ of the determinant of the corresponding normal equations. In analogy to the preceding section we assign a relative weight $W(p_o, r_o)$ to each of the digital filters (4.1) as follows:

$$W(p_o, r_o) = \frac{E(p_o, r_o)}{\left[\sigma(p_o, r_o)\right]^{p_o + r_o + 1}}. \tag{4.2}$$

Let (p_1, r_1), (p_2, r_2) and (p_3, r_3) be the pairs of (p_o, r_o) corresponding to the three largest values of $W(p_o, r_o)$. For $s = 1,2,3$ let

$$\bar{x}_j(p_s, r_s) = \hat{x}(j),$$

if $n_o + 1 \leqslant j \leqslant N - n_o$. Let

$$\bar{x}_j(p_s, r_s) = x(j),$$

if $j > N - n_o$. Let

$$\hat{x}_j'(p_s, r_s) = \hat{x}'(j),$$

if $n_o + 1 \leqslant j \leqslant N - n_o$. Let

$$\hat{x}'_j(p_s, r_s) = \sum_{i=o}^{p_s} d_i(p_s, r_s)\hat{x}(j - iq_o) + \sum_{i=1}^{r_s} f_i(p_s, r_s)\hat{x}'(j - iq_o),$$

if $N \geqslant j > N - n_o$. Then the estimates of the first derivative in the final interval of the data are given by

$$\hat{x}'(j) = \sum_{s=1}^{3} W(p_s, r_s)\hat{x}_j'(p_s, r_s) \Big/ \sum_{s=1}^{3} W(p_s, r_s) \tag{4.3}$$

$$j = N - n_o + 1, N - n_o + 2, \ldots, N.$$

The estimates of the second derivative are obtained in the same manner by replacing $\hat{x}'(j)$ by $\hat{x}''(j)$ for $j = n_o + 1, n_o + 2, \cdots, N - n_o$ in (4.1) and then by solving the resulting overdetermined system, i.e., by determining the corresponding linear digital filters. These are used to obtain the estimates of the second derivative analogous to (4.3).

The estimates of the derivatives in the initial segment of the data are obtained by renumbering the data points as follows: $\xi(j) = x(N - j + 1)$ and then by applying the procedure just described to $\xi(j)$.

5. NUMERICAL EXAMPLES

We compute approximations of the first and second derivative by the method described on the preceding pages for ten sets of synthetic data. The ten cases numbered in Column 1 of Table 1 are obtained by evaluating the functions $x(t)$ listed in Column 2 in the intervals given in Column 3 at the step size Δ shown in Column 4 and then by adding to each value a pseudo-random error normally distributed with zero mean and standard deviation σ as contained in Column 5. In each case the standard deviation σ is equal to the average of the absolute values of the change of the corresponding function as its argument changes by Δ. With random errors of this size derivatives cannot be estimated by divided differences.

In Case 9 $\sigma = 0$, i.e., here we have functional values exact up to 15 decimal digits, which is single precision for the computer employed.

In Case 5 the function $x(t) = J_0(t)$ is the Bessel function of the first kind. In Case 7 (Witch of Agnesi) $x(t) = 2\cos^2(\tan^{-1}\frac{t}{2})$ and in Case 8 we have a rational function $x(t) = 36t/(t^2 + 9)$. These three functions as well as the error function are included to test how well their derivatives can be approximated when the functions do not belong to the Algebra A defined in Section 2.

TABLE 1. SYNTHETIC DATA

1	2	3	4	5
Case	$x(t)$	Interval	Δ	σ
1	$\sin 2\pi t$	[0,1]	.004	.008
2	$\sin 2\pi t + 0.1 \sin 10\pi t$	[0,1]	.004	.022
3	e^t	[0,5]	.01	.295
4	$4^t (1 + t + t^2 + t^3)$	[0,2]	.004	.478
5	$J_0(t)$	[1,6]	.01	.00344
6	error function	[0,2]	.005	.00119
7	Witch of Agnesi	[0,2]	.005	.0025
8	Newton Serpentine	[0,2]	.005	.0139
9	$2t^3 - 9t^2 + 12t$	[0,3]	.005	0
10	$2t^3 - 9t^2 + 12t$	[0,3]	.005	.01833

The results are summarized in Table 2. Here Column 1 is the same as in Table 1. To each row of Column 1 there correspond two rows of the remaining columns. One of these contains the errors of the first derivative $\hat{x}'(t)$ and the other that of the second derivative $\hat{x}''(t)$ as indicated in Column 2. Columns 3 and 4 give RMSE's of the respective estimates of derivatives expressed in percent of the RMS of the actual values at the data points. Column 3 (B.S.) shows the errors of the estimates obtained by fitting the noisy data by B-splines. Column 4 contains the errors (in percent) of the estimates computed by the current method (C.M.), i.e., by the method of this report. By comparing these two columns we see that the method described here gives much better estimates of derivatives in all but three cases out of 20. Two of these cases are for exact values of the cubic polynomial. The first and the second derivatives computed by cubic B-spline approximation agree up to eight digits with the exact values in this case. The coefficients of the autoregressive model with highest weight agree with the exact coefficients up to seven digits. The resulting perturbation of one or two units in the eighth digit yields eigenvalues accurate up to only three decimal digits. The derivatives computed with these approximations on the average differ from the exact values in the fourth decimal digit, as shown in Column 4. The last two rows show that the method of this report produces considerably better results than the B-spline approximation when the values of this same polynomial contain random errors.

In Case 2 the error for the first derivative obtained by the method of this report is almost twice as big as that obtained by the spline approximation. However, the error of the second derivative computed by the current method is only half the error obtained by the spline.

Of the remaining sixteen derivatives (aside from Cases 2 and 9) five approximations by the current method are better by an order of magnitude or more than that by the spline; six approximations have errors four or more times smaller than the errors by the spline method, and the remaining five are at least twice as good as those by the spline procedure. The worst error by the spline approximation as compared to the approximation by the current method is that of the second derivative of the exponential function. Even if this worst case is ignored, the error by the spline approximation is on the average 474% of the error by the method of this report.

Compatibility of derivatives computed by our method with the data was further tested by comparing exact functional values with an approximate first integral of the first derivative and the second integral of the second derivative. Approximations of the integrals were obtained by trapezoidal rule. The RMSE of the integral of the first derivative computed for various segments of the central portion of the data varies from 0.075 to 1.88 percent of the RMS of the functional values. The smallest error is for Witch of Agnesi (Case 7) and the largest for Case 4. The values of the derivatives in the initial and final data segments, however, are less accurate. Consequently, the RMSE's of the integrals of the first derivative computed for various segments of the complete data set vary from 0.10 (Case 3) to 3.8 (Case 5) percent. Similarly, the RMSE's of the second integral of the second derivative in the central portion of the data vary from 0.26 (Case 3) to 17.5 (Case 4) percent and in the complete set of data between 0.25 (Case 3) and 43.5 (Case 2) percent.

TABLE 2. APPROXIMATION ERRORS IN PERCENT

1	2	3	4
Case	Deriv	B.S.	C.M.
1	x'	2.5	.61
	x''	20	1.1
2	x'	3.3	6.2
	x''	23	11.9
3	x'	3.3	.22
	x''	67	.25
4	x'	3.2	1.6
	x''	67	8.6
5	x'	4.7	2.2
	x''	140	15.5
6	x'	4.5	1.7
	x''	253	37
7	x'	3.8	1.9
	x''	191	36
8	x'	4.7	1.1
	x''	347	28
9	x'	0	.042
	x''	0	.071
10	x'	4.29	1.55
	x''	88	3.55

We note further that our method is not too time-consuming in spite of the fact that it determines several approximations and the corresponding computer program contains several loops. Computing time on the CDC CYBER 76 computer of BRL, including generation of synthetic data, in none of the ten cases exceeded 8.8 sec. The average computing time per case was 4.5 sec.

6. EXTRAPOLATION/FORECASTING

A variant of the new method described in this and earlier related reports can be used for extrapolation out of the region in which the data was sampled. As with any forecasting procedure it must be assumed that the underlying trend shown by the sample continues into the region considered. If conditions change that trend significantly, the forecast will naturally be much less reliable.

If equation (2.14) remains valid, functional values $s \cdot q \cdot \Delta t$ later in time are obtained by replacing n by $n + s$, while the eigenvalues λ remain the same. Other equations are modified accordingly.

In the central portion of the sample a span of $2k + 1$ points, $q \cdot \Delta t$ apart, was used to find values at a point t offset zero from the center of the span. In the right-hand end portion the offset was changed to $k \cdot q \cdot \Delta t$ so as to use available data points, to the left of and including t, in the evaluation. If the offset is increased further, to $v \cdot q \cdot \Delta t$ for some $v > k$, the span can remain within the data sample even though t is outside. This provides the basis of forecasting.

The correspondingly modified algorithm was used for extrapolation of several of the standard test functions. In cases where the λ's closely approximate those of the exact (non-noisy) function, extrapolated values well outside the sample were obtained with acceptable accuracy. In cases of greater noise, poorer approximation of the true λ's is typical, and extrapolation is therefore less satisfactory.

For $x = \sin 2\pi t + \epsilon$, 201 points on [0,1], prediction of the extrema and zeroes on [1,2] was undertaken. For a noise level ($\sigma(\epsilon)$) of .001 the zeroes at $t = 1.5$ and 2.0 were predicted to be at $t \pm 1 \cdot \Delta t$ with error in $\hat{x}(t)$ less than .001 in those intervals. Similarly, the extrema at 1.25 and 1.75 were predicted at $t \pm 1 \cdot \Delta t$ with $|\hat{x}|$ values on (0.999, 1.000). The corresponding first derivatives $\hat{x}'$, computed by the linear filtering method of Section 4, were found with RMS error within 0.2 percent of the overall RMS (x').

Corresponding results with greater noise ($\sigma = .010$) also located the extrema and zeroes within $1 \cdot \Delta t$, with $|\hat{x}|$ in error by .010 (RMS error) over the entire interval [1, 2] The RMS error for $\hat{x}'$ in the extrapolated region was 3.0% of RMS (x').

For the more oscillatory function $x = \sin 2\pi t + 0.1 \sin 10\pi t + \epsilon$, with $\sigma(\epsilon) = .001$, the extrema at $t = 1.25$ and 1.75 and the zeroes at $t = 1.5$ and 2.0 were predicted to within $1 \cdot \Delta t$, and the values near the extrema were correct within .002. The corresponding first derivatives $\hat{x}'$ near the zero at $t = 1.5$ were in error by less than 3% of true value, degrading somewhat thereafter. The corresponding case with $\sigma = .010$

extrapolated $\hat{x}$ well (RMSE of .02 on [1, 2]) but $\hat{x}'$ began to degrade somewhat earlier in the extrapolation region. Note that the increased input noise is a significant fraction of the amplitude of the second term of this double sine function.

Extrapolation should always be used with caution, but there are many practical cases where it may be necessary. Economic forecasting, demographic projections, and prediction of future locations of moving targets are examples. The reverse process of extrapolating into the past is also of interest, e.g., likely earlier values of a variable with only recent measurements available. This can be achieved by storing the data values in reverse and changing the sign of first, third, and other odd derivatives of interest. The current generalized algorithm thus can be used in either context.

REFERENCES

1. C. Masaitis and G. Francis, "Numerical Differentiation of Noisy Data," ARBRL-MR-03126, Aug 81, AD A104631.

2. C. Masaitis and G. Francis, "Computation of Derivatives from Tabular Data," ARBRL-MR-03188, Jul 82, AD B066124L.

APPENDIX

DNDLIB - A USER LIBRARY FOR DIFFERENTIATING NOISY DATA

There now exists on MFZ, the BRL CDC Cyber 7600 computer, a user library called DNDLIB, created by G. Francis in September 1981 and based on the computer programs and theoretical algorithms discussed in three BRL publications by C. Masaitis and G. Francis, this report and its earlier companions.

The library was designed for easy access by other users of the BRL computer. In general only one of the 14 subroutines included need be called by a user program, as the others are called as needed automatically. That particular subroutine is named DND and allows a certain amount of flexibility on the part of the user. He must set up an array (vector) of equally spaced data points (usually experimental values containing noise, i.e., measurement errors and the like) and specify the number of such points, say N. In addition he must select the maximum order of derivatives of interest (1 to 3) and provide additional arrays for smoothed results and each order of derivative. The current upper bound on N is 1000, and N should exceed 30.

A sample call (in Fortran) is as follows:

```
CALL DND (XN, N, DT, MXOD, XS, D1, D2, D3)
```

where, say, N = 250, XN is an array of at least N values (points 1 to N to be processed), DT is the spacing of the independent variable, MXOD is the maximum order of derivative wanted (1 to 3), and arrays XS (of length N or more) and D1 (similarly) are to receive the smoothed values and the first derivatives at corresponding points. If MXOD is 2 or more, an array D2 must be provided; likewise if MXOD is 3, an array D3 is needed. The labels D2 and D3 may be omitted from the call if MXOD is 1.

For each point i, i=1 to N, there will be a smoothed value at XS(i) and a first derivative at D1(i). If MXOD is 2 a numerical second derivative will be at D2(i), and if MXOD is 3 then D3(i) will contain the third derivative corresponding to point i. (If still higher derivatives are desired, the process can be repeated by copying D3, say, into XM and making a new call. The accuracy of derivatives grows worse, however, as the order increases, particularly if the original data contains considerable error.)

The method of DNDLIB uses a few subroutines from the IMSL library, so that library must be made available, too. This is done on MFZ by means of the following CDC 'control cards' or suitable replacements:

```
ATTACH(IMSL)
ATTACH (DNDLIB, ID=PUBLIC)
LIBRARY(IMSL,DNDLIB)
```

In addition the standard BRL subroutines known as FNEQS and MATINV are used for matrix setup and inversion. These are available on MFZ with no user action required.

The size of the full set of subroutines is approximately 20K words of SCM (central memory) and 12K of LCM plus the user-supplied arrays (of 3N to 5N words, as mentioned earlier). This includes all work space for N up to 1000 and MXOD up to 3, mostly in LCM.

The time required for calculations is highly dependent on the number of data points and also on the order of derivatives wanted as well as the degree of complexity of the underlying function. For a sample noisy sinusoid with derivatives wanted through order 2 the following times were found (in cpu seconds on MFZ):

N =	100	250	500	1000
T =	0.5	1.3	4.4	11.1

As indicated above results are found at all N points, not just a few.

If extrapolation is required, an alternate call of much the same form is used, with one additional parameter, denoted NP, and the subroutine name is DNDE rather than DND:

```
CALL DNDE (XN, N, DT, MXOD, NP, XS, D1, D2, D3)
```

A request for extrapolation forward is entered by specifying NP greater than N, but under 1000. This feature should be used with caution. Naturally, all arrays for results should be of at least NP cells.

It is hoped that this library will prove of benefit to users, who are encouraged to apply it to their own noisy data, especially when the assumptions of Section 2 are thought to be satisfied.

DISTRIBUTION LIST

No. of Copies	Organization
12	Administrator Defense Documentation Center ATTN: DTIC-DDA Cameron Station Alexandria, VA 22314
1	Director Defense Advanced Research Projects Agency ATTN: Mr. Charles Lehner 1400 Wilson Blvd. Arlington, VA 22209
1	Deputy Assistant Secretary of the Army for R&D WASH, DC 20310
1	HQDA (DAMA-ZD) ATTN: Mr. H. Woodall WASH, DC 20310
4	HQDA (DAMA-AOA-M, LTC H. G. Glock DAMA-RA, MAJ Yeats DAMA-WSW-A, MAJ Scott DAMA-CSS, LTC R. Bertils) WASH, DC 20310
2	HQDA (DAMO-ROD, MAJ Halbert DAMO-TOD, MAJ Miller) WASH, DC 20310
1	HQDA (DACA-CAM, LTC Toner) WASH, DC 20310
1	Commander in Chief, USAREUR & 7A ATTN: AEAGB-PDN-SE APO New York 09403
1	Commander US Army Concepts Analysis Agency ATTN: MOCA-FDS, LTC G.W. Hausler 8120 Woodmont Avenue Bethesda, MD 20014
1	Commander US Operational Test and Evaluation Agency ATTN: DACS-TEM-G, COL Hauser Ft. Belvoir, VA 22060
1	Commander US Army Materiel Development and Readiness Command ATTN: DRCDMD-ST 5001 Eisenhower Avenue Alexandria, VA 22333
1	Director US Army Armament Rsch & Dev Cmd Benet Weapons Laboratory ATTN: DRDAR-LCB-TL Watervliet, NY 12189
1	Commander US Army Materiel Development and Readiness Command ATTN: DRCPM-GCM-WF LTC Peter McVey 5001 Eisenhower Avenue Alexandria, VA 22333
1	Commander US Army Materiel Development and Readiness Command ATTN: DRCLDC, Mr. T. Shirata 5001 Eisenhower Avenue Alexandria, VA 22333
2	Commander USAARRADCOM ATTN: DRDAR-TSS Dover, NJ 07801
6	Commander USAARRADCOM ATTN: DRDAR-CG DRDAR-LCS, Mr. Killen DRDAR-SCF, Mr. Pfleger DRDAR-SCF-C, Mr. Baumgarthe DRDAR-LCS-A,Mr. Hauptman DRDAR-TDC, Dr. D. Gyorog Dover, NJ 07801

DISTRIBUTION LIST

No. of Copies	Organization
1	Commander US Army Armament Materiel Readiness Command ATTN: DRSAR-LEP-L Rock Island, IL 61299
1	Commander US Army Aviation Research and Development Command ATTN: DRDAV-E 4300 Goodfellow Blvd. St. Louis, MO 63120
1	Director US Army Air Mobility Research and Development Laboratory Ames Research Center Moffett Field, CA 94035
1	Commander US Army Communications Research and Development Command ATTN: DRDCO-PPA-SA Fort Monmouth, NJ 07703
1	Commander US Army Electronics Research and Development Command Technical Support Activity ATTN: DELSD-L Fort Monmouth, NJ 07703
1	Commander US Army Harry Diamond Labs ATTN: DELHD-TA-L 2800 Powder Mill Road Adelphi, MD 20783
1	Commander US Army Missile Research and Development Command ATTN: DRSMI-R Redstone Arsenal, AL 35898
1	Commander US Army Missile Research and Development Command ATTN: DRSMI-YDL Redstone Arsenal, AL 35898
2	Commander US Army Mobility Equipment Research & Development Cmd ATTN: DRDME-WC DRDME-RZT Fort Belvoir, VA 22060
1	Commander US Army Tank Automotive Research and Development Command ATTN: DRSTA-TSL Warren, MI 48090
4	Commander US Army Tank Automotive Research and Development Command ATTN: DRDTA-TWL, Dr. J. Parks DRDTA-R, Mr. C. Bradley DRDTA-WG, Mr. Slawski DRDTA-R, Mr. V. Pagano Warren, MI 48090
4	Commander US Army Tank Automotive Research and Development Command ATTN: DRDTA-RH, Mr. Ted Puree, Mr. Ron Beck DRDTA-R M551 PIP Office Warren, MI 48090
5	President US Army Armor & Engineer Board ATTN: STEBB-TA COL Boehme LTC John Dennis COL Fitzmorris Mr. William Hahn Fort Knox, KY 40121
2	Project Manager, XMI ATTN: MG R.J. Baer MAJ A. Lipinski 28150 Dequindre Street Warren, MI 48092
1	Commandant US Army Infantry School ATTN: ATSH-CD-CSO-02 Ft. Benning, GA 31905

DISTRIBUTION LIST

No. of Copies	Organization
1	Project Manager IFV 28150 Dequindre Street Warren, MI 48092
4	Commander US Army Training & Doctrine Command ATTN: Classified Library ATCD-CM-AR, COL Kelly ATCD-A, BG Clarke COL James Broden Fort Monroe, VA 23651
2	Commander US Army TRADOC Systems Analysis Activity ATTN: ATAA-SL ATAA-T, Mr. Shook White Sands Missile Range, NM 88002
2	Commander US Army Combined Arms Combat Developments Activity ATTN: ATCACC-MA, MAJ Smith ATCA-CAT-D, MAJ Lenz Fort Leavenworth, KS 66027
1	Commander US Army Logistics Center ATTN: ATCL-MT, CW2 McFadden Fort Lee, VA 23801
2	Commander US Army Armor Center & School ATTN: ATZK-CD, COL Pigg ATZK-DCG Fort Knox, KY 40121
3	Commandant US Army Armor School ATTN: ATSB-CD-MM, COL Demonte CPT White ATSB-C, BG Williams Fort Knox, KY 40121
1	Commander 194th Armored Brigade Fort Knox, KY 40121
1	Commander US Army Field Artillery School ATTN: Combat & Training Div MAJ Michela Fort Sill, OK 73503
1	Commander US Army Infantry School ATTN: ATSH-MV Fort Benning, GA 31905
1	Commander Headquarters, TCATA Fort Hood TX 76544
1	Commander David W. Taylor Naval Ship Research and Development Center Bethesda, MD 20084
1	Commander Naval Surface Weapons Center Silver Spring MD 20910
2	Superintendent Naval Postgraduate School ATTN: Dr. Sam H. Parry Dr. Russel Stolfi Monterey, CA 93940
1	General Accounting Office Procurement and Systems Acquisition Division ATTN: Mr. Richard Tuck 441 G Street, NW WASH, DC 20548

DISTRIBUTION LIST

No. of Copies	Organization
1	AFELM, The Rand Corporation ATTN: Library-D 1700 Main Street Santa Monica, CA 90406
1	Director Battelle Advanced Concept Lab Columbus Division ATTN: Dr. Robert Stein Warren, MI 48090

No. of Copies	Organization
	Aberdeen Proving Ground
	Cdr, USATECOM ATTN: DRSTE-TO-F
	Dir, USAMSAA ATTN: DRXSY-D DRXSY-MP, H. Cohen Mr. J. Zeller Mr. D. Woomert
	Dir, USAHEL ATTN: Mr. A. J. Eckles, III
	Dir, USACSL, Bldg. E3516, EA ATTN: DRDAR-CLB-PA DRDAR-CLN DRDAR-CLJ-L

USER EVALUATION OF REPORT

Please take a few minutes to answer the questions below; tear out this sheet, fold as indicated, staple or tape closed, and place in the mail. Your comments will provide us with information for improving future reports.

1. BRL Report Number _____

2. Does this report satisfy a need? (Comment on purpose, related project, or other area of interest for which report will be used.)

_____

_____

_____

3. How, specifically, is the report being used? (Information source, design data or procedure, management procedure, source of ideas, etc.) _____

_____

_____

4. Has the information in this report led to any quantitative savings as far as man-hours/contract dollars saved, operating costs avoided, efficiencies achieved, etc.? If so, please elaborate.

_____

_____

5. General Comments (Indicate what you think should be changed to make this report and future reports of this type more responsive to your needs, more usable, improve readability, etc.) _____

_____

_____

_____

6. If you would like to be contacted by the personnel who prepared this report to raise specific questions or discuss the topic, please fill in the following information.

Name: _____

Telephone Number: _____

Organization Address: _____

_____

_____

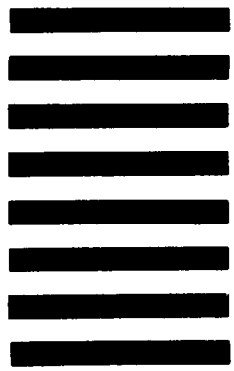
FOLD HERE

Director
US Army Ballistic Research Laboratory
ATTN: DRDAR-BLA-S
Aberdeen Proving Ground, MD 21005

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

BUSINESS REPLY MAIL
FIRST CLASS PERMIT NO 12062 WASHINGTON, DC
POSTAGE WILL BE PAID BY DEPARTMENT OF THE ARMY



Director
US Army Ballistic Research Laboratory
ATTN: DRDAR-BLA-S
Aberdeen Proving Ground, MD 21005

FOLD HERE